Development of a Grape Genomics Database Using IBM DB2 Content Manager Software

Hong Huang[1,4,a], Jiang Lu[1,b], W. Hunter[2,c], S. Dowd[3], C. Katsar[2] and C. Jörgenson[4]
[1]Center for Viticulture and Small Fruit Research, Florida A&M University, Tallahassee, FL, 32308, USA
[2]United States Department of Agriculture, Agriculture Research Service, United States Horticultural Research Laboratory, Fort Pierce, Florida, 34945, USA
[3]United States Department of Agriculture, Agriculture Research Service, Livestock Issues Research Unit, Lubbock, TX, 79403, USA
[4]College of Information, Florida State University, Tallahassee, FL, 32304, USA

Keywords: Pierce's disease, *Vitis*, Expressed Sequence Tags (ESTs), functional annotation, ontology, chromatogram, XML, pathway

Abstract

Diseases are a limiting factor for the growth non-native grapes in many areas of the United States. For example, Pierce's disease limits the production of European grapes (*Vitis vinifera*) in most part of the southeastern United States. Disease resistance of the non-native grape could be enhanced by incorporation of the resistance traits of native American grape species such as *Vitis shuttleworthii*, *V. aestivalis*, *V. riparia*, and *V. rotundifolia* through the use of cross pollination. Alternatively, a 'molecular breeding approach' could be used to assist and accelerate the conventional breeding process. The recent development and availability of large number of gene sequence data makes the molecular approach more realistic than ever before. Florida A&M University (FAMU) has developed a grape genomics database using IBM DB2 Content Manager to store genetic sequences such as Expressed Sequence Tags (EST), genomic sequences, molecular markers, chromatograms, graphic information, blast hits, and functional annotations in an organized and coherent fashion. The system provides a convenient way to associate heterogeneous data sets such as genetic sequences, and various annotations with extensive search capabilities. This database system will be better served for the biological data integration and information retrieval relative to *Vitis* spp.

INTRODUCTION

Grapes are among the most important fruit crops in the world. In the United States of America, more than 40 states have commercial vineyards. Pierce's Disease (PD) is caused by *Xylella fastidiosa*, and is a major limiting factor in production of bunch grape in the southeastern United States. The grape industry in the southeastern US is therefore based primarily on some PD-resistant *Vitis* species that are native to the Gulf Coastal regions of the US. PD-resistant species from this region include the muscadine grape (*Muscadinia* Planch.), *V. aestivalis*, *V. shuttleworthii*, and their interspecific hybrids.

Expressed Sequence Tags (ESTs) are relatively short cDNA sequences and are representative samples of the transcribed portion of the genome being studied (Adams et al., 1991; Driesel and Lommele, 2003). ESTs from the European grape (*V. vinifera)* have been thoroughly investigated. The International Grape Genomics community has been conducting studies on grape transcriptional expression analysis for stress and defense relative gene identification for years (Ablett et al., 2000; da Silva et al., 2004). However, the native American grape species such as *V. shuttleworthii*, *V. aestivalis*, and *V. rotundifolia* (muscadine), which is highly resistant to various diseases such as Pierce's

[a] huanghon2003@yahoo.com
[b] jiang.lu@famu.edu
[c] WHunter@ushrl.ars.usda.gov

Proc. IX[th] Intl. Conf. on Grape Genetics and Breeding
Eds.: E. Peterlunger et al.
Acta Hort. 827, ISHS 2009

Disease, anthracnose, black rot, downy mildew, and powdery mildew diseases have not been investigated until recently (Lu and Hunter, 2004). It is logical to collect all the *V. shuttleworthii* and muscadine ESTs along with other public available ESTs from native American grape species such as *V. aestivalis* and *V. riparia* to compile into a single set of genomics resources containing rich disease resistant genes candidates stored in the database.

The database community has manifested this interest in heterogeneous data management and efficient retrieval of iconic information by content (Brown and Jurisica, 2005; Poustelnikova et al., 2004). Our goal is to build a user-friendly, searchable database for the research and academy community for the retrieval and visualization of grape genomic data.

To achieve this, we developed a database including all available native American grape genomics data to be stored, shared, searched, and aided in data annotation, gene/protein discovery, and/or characterization. IBM DB2 Content Manager provides a scalable repository system for the capture, creation, organization, and retrieval of content.

MATERIALS AND METHODS

Data Source Collection and Integration

ESTs, including 16,000 from *V. shuttleworthii* and 7,000 from muscadine developed by FAMU-USDA collaboration were collect along with retrieved data sets of 2,177 from *V. aestivalis*, and 1,995 *V. riparia* EST sequences available in NCBI databases as the data sources for database development.

ESTs Functional Annotation and Ontology/KEGG Pathway Classification

The EST sequences were functional annotated by HT-GO-FAT program (Dowd and Zaragosa, 2005). HT-GO-FAT is a software data mining and annotation toolkit developed in Microsoft .NET 2003 development environment. HT-GO-FAT was utilized to functionally annotate the assembled sequences with their associated Gene Ontology, Enzyme Commission numbers, and their associated KEGG mappings. The annotation results were stored in the XML format for database integration.

IBM DB2 Content Manager System Configuration and Architecture

Content Manager uses triangular architecture, with Library Server and Resource Manager as the foundation, and client applications on top. There are various client options (Client for Windows, eClient, and customized client) available. Content Manager system architecture can be built with a two-tier or three-tier configuration using different client options, on individual or mixed platforms (Fig. 1). It provides workflow routing, life cycle management as well as content sharing. The triangular architecture in Content Manager, which was installed on top of Linux Suse operation system, contains a Library Server, Resource Manager, and Content Manager Clients for a three-way communication.

RESULTS AND DISCUSSION

Technical Summary of FAMU "Vitigene" Database

The system integrate Content Manager has been installed on top of Linux Suse operation system. The Resource Manager and Eclient were configured to provide a unified storage/retrieval and importing and exporting data solution that can eventually include features for automatic data processing, avoiding the many manual steps that are currently needed. The grape genomics database is being constructed to conveniently link different item types among the heterogeneous data sets including genetic sequences, chromatograms, blast results, and gene ontology within the same biological domain in a parent-child hierarchy structure (Fig. 2). The database also provides a convenient index and full text searching (Figs. 3 and 4). The functionally annotated EST sequences were imported into the "Vitigene" database with wildcard search and full text search

capabilities. The specific EST sequences and their functional annotations information including sequence identifier, blast definition, gene ontology, KEGG pathway, and EC numbers can be exported and downloaded from the database system. Such a system easily deploys the pictures referencing text data using cross link and folder management. The system has following advantages - heterogeneity: multiple format files integration; data sharing and retrieval: full text searches, wildcard searches and semantic searches; automated update tools: user-friendly data modeling interface; users' authentication and authorization; a content-based management system including document routing; and workflow management.

The database is expanding to host the function and genetic and/or physical map location of grape critical genes as well as identified SSR and SNP markers linked to agronomically important phenotypes. In this way, the database helps to expedite overall grape research and improvement efforts. The system can provide data retrieval and storage service for the grape researchers, biologists, bioinformatics specialists, information scientists, and computational scientists around the world. It also has the capacity to cross-across database for wider use and recognition. This database system can cross link and index heterogeneous data and thus provides a convenient way to build up an efficient database for hosting genomic datasets. The development of such a database opens new doors in bioinformatics and promotes the knowledge sharing and integration in both the industrial and academic worlds. The system described here provides a convenient platform to train scholars and students in database management either on-site or via remote access.

ACKNOWLEDGEMENTS

This work was supported by USDA-FAMU science center program.

Literature Cited

Ablett, E., Seaton, G., Scott, K., Shelton, D., Graham, M.W., Baverstock, P., Lee, L.S. and Henry, R. 2000. Analysis of grape ESTs: global gene expression patterns in leaf and berry. Plant Science 159:87-95.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., W.A., Olde, B. and Moreno, R.F. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656.

Brown, K.R. and Jurisica, I. 2005. Online Predicted Human Interaction Database. Bioinformatics 2:2076-2082.

da Silva, F.D., Landolino, A.B., Lim, H.J., Choi, H.K., Baek, J., Leslie, A.D., Xu, J. and Cook, D.R. 2004. Identification of transcripts correlated with berry development and host responses to Pierce's disease of grapes. P41. Plant Animal Genome Conference XII. San Diego, CA.

Dowd, S.E. and Zaragoza, J. 2005. High Throughput Gene Ontology Functional Annotation Toolkit (Ht-Go-Fat) Utilized for Animal and Plant. Plant and Animal Genome XIII Conference, San Diego, CA.

Driesel, A.J. and Lommele, A. 2003. Towards the transcriptome of Grapevine (*Vitis Vinifera* L.) Plant & Animal Genomes XI Conference. San Diego, CA.

Lu, J., Hunter, W., Dang, P., Huang, H. and Leong S. 2004. Identification of Disease Defense- and Stress –Related Genes in the Grape *Vitis shuttleworthii* though EST analysis. Plant and Animal Genome XII Conference, San Diego, CA.

Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M. and Reinitz, J.A. 2004. Database for management of gene expression data *in situ*. Bioinformatics 20:2212-2221.

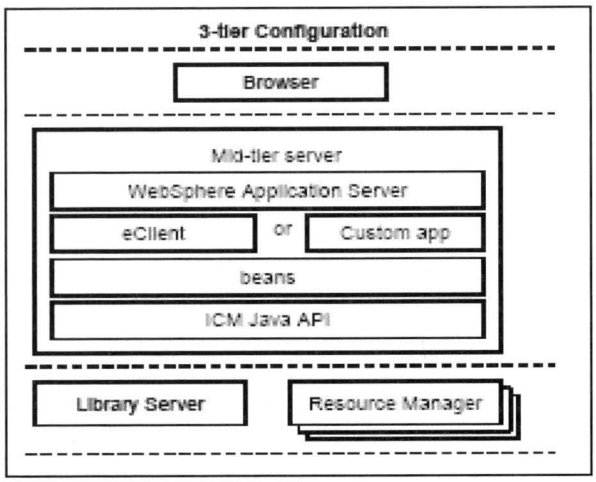

Fig. 1. Content manager 3 tier configuration.

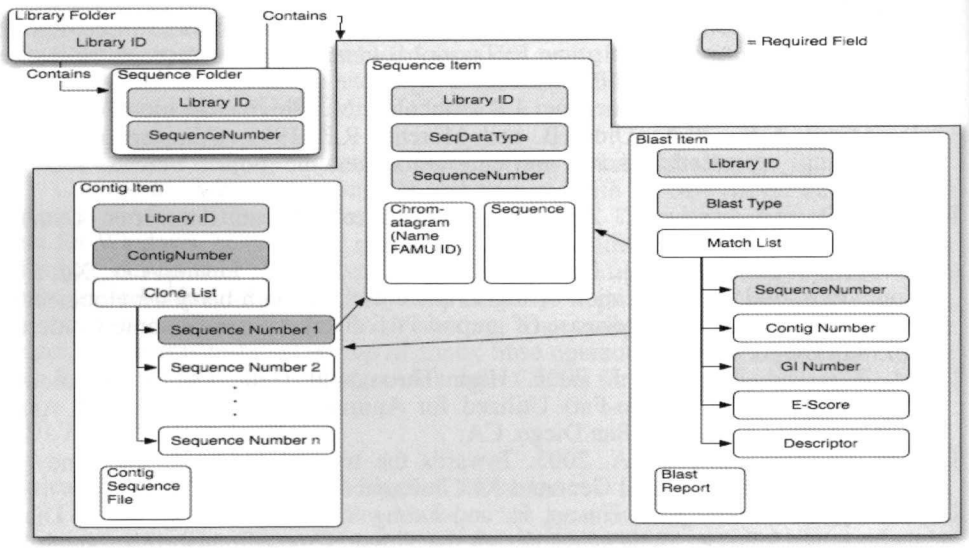

Fig. 2. The current configured data model for Content Manager.

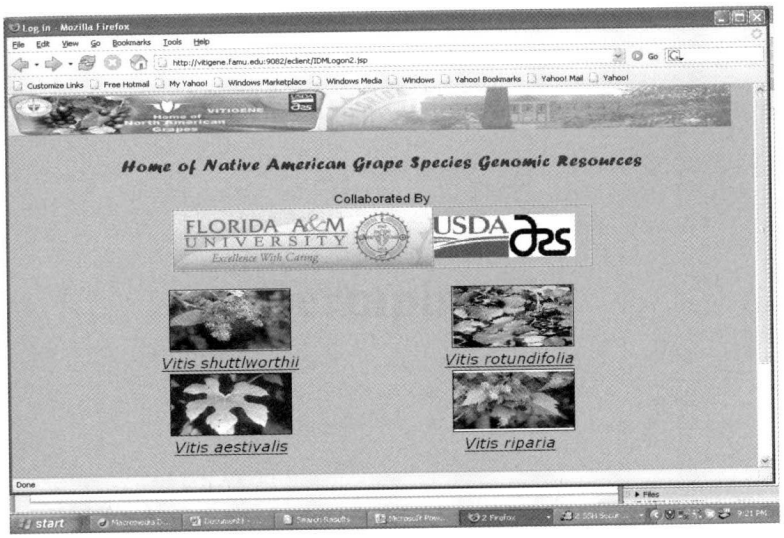

Fig. 3. The front page included the native American grape species database.

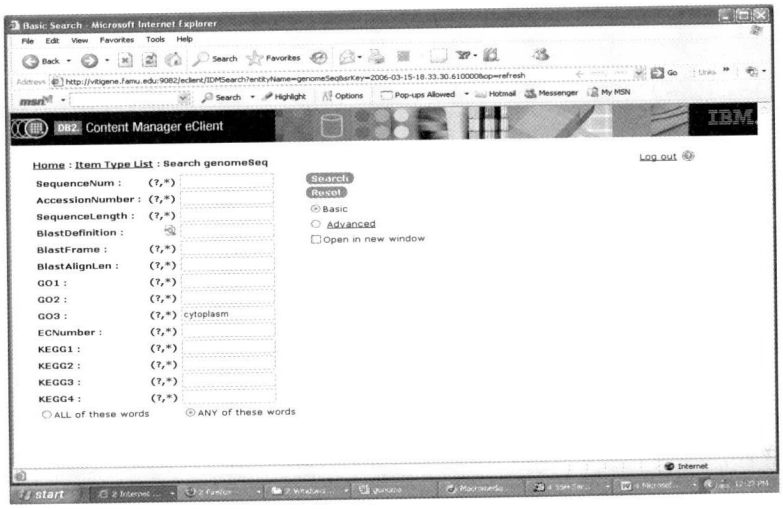

Fig. 4. A wildcard search example: to retrieved sequences containing "*cytoplasm*" in blast definition.